Sound

The following chapter introduces another basic perceptual sensation: Sound. Hearing together with vision form the two most important sensual inputs that a human can process. Being deaf is considered a serious disability as is, of course, being blind. Even though man parallels exist between visual signal processing and acoustic signal processing, sound has unique properties -- often orthogonal to those of visual signals. One could say this is the reason nature chose to give animals both visual and acoustic sensors: To gather complementary information about the environment. Sound is used by many species for detecting danger, for navigation, predation, and communication. Earth's atmosphere, water, and virtually any physical phenomenon, such as fire, rain, wind, surf, or earthquake, produces unique sounds. Many species, such as frogs, birds, marine and terrestrial mammals, have also developed special organs to produce sound. In some species, these have evolved to produce singing and speech. Furthermore, humans have developed culture and technology (such as music, telephone, and radio) that allows them to generate, record, transmit, and broadcast sound. This is of course the reason why you read about sound in this book. The following chapter introduces the basic properties of sound, sound production, and sound perception. Of course, inside the scope of a multimedia book we can only scratch on the surface of a very complex and fascinating topic.

What is sound?

The American Heritage Dictionary of the English Language, Fourth Edition defines sound as “a traveling wave which is an oscillation of pressure transmitted through a solid, liquid, or gas, composed of frequencies within the range of hearing and of a level sufficiently strong to be heard, or the sensation stimulated in organs of hearing by such vibrations.”

This very compressed formulation gives us a perfect start to discuss the properties of sound. Sound is generated by any mechanic oscillation. Despite light, sound is traveling through a medium. In a a vacuum, for example, there is no sound and, of course, one can’t hear exploding space ships. The traveling speed of sound is varied according to the medium sound travels in: In dry air at 20 °C, the speed of sound is 343 meters per second or Mach 1. In fresh water, also at 20 °C, the speed of sound is approximately 1,482 m/s.

For a sound to be heard, the frequency and the amplitude of the oscillation has to be in a certain range. For humans, hearing is normally limited to frequencies between about 12 Hz and 20,000 Hz (20 kHz). The upper limit generally decreases with age. Other species have a different range of hearing. Dogs, for example, can perceive vibrations higher than 20 kHz. This is also one of the reasons why dogs and cats will not react to broadcast TV the same way as humans do. While a

mouse can be draw the attention of a cat from hundreds of meters away, the same sound, much more intense, from and MP3 player or a TV might not be interesting at all: Multimedia compression and signal processing assumes human perception. Producing the equivalent of a TV for a dog or a cat is a completely different story.

Since sound is an oscillation of pressure, the amplitude of a sound wave can be measured by measuring the so-called sound pressure. Sound pressure is defined as the difference between the average local pressure of the medium outside of the sound wave in which it is traveling through (at a given point and a given time) and the pressure found within the sound wave itself within the medium. The square of this difference is usually averaged over time and/or space, and a square root of such average is taken to obtain a root mean square (RMS) value.

As the sound pressure perceived by the human ear is non-linear (see Chapter XXX [mulaw]) and the range of amplitudes is rather wide, sound pressure is often measured as a level on a logarithmic scale, the so-called decibel scale. The sound pressure level (SPL) or L_{p} is defined as:

$$L_{\mathrm{p}} = 10\ \log_{10}\left(\frac{p^2}{{p_{\mathrm{ref}}}^2}\right) = 20\ \log_{10}\left(\frac{p}{p_{\mathrm{ref}}}\right)\ \mathrm{dB}$$

where p is the root-mean-square sound pressure and p_{ref} is a reference sound pressure. Commonly used reference sound pressures for silence, defined in the standard ANSI S1.1-1994, are 20 µPa in air and 1 µPa in water. Most sound recording equipment are calibrated to omit 0-amplitude at these levels.

As will be explained later in this chapter, the human ear does not have a flat spectral response. i.e. the same sound pressure at a different frequency will be perceived as different volume levels. Therefore, sound pressures are often frequency weighted so that the measured level will match perceived levels more closely. The most common one is the so-called A-weighting scheme, defined by IEC. Figure 1 shows a graph of the scheme.

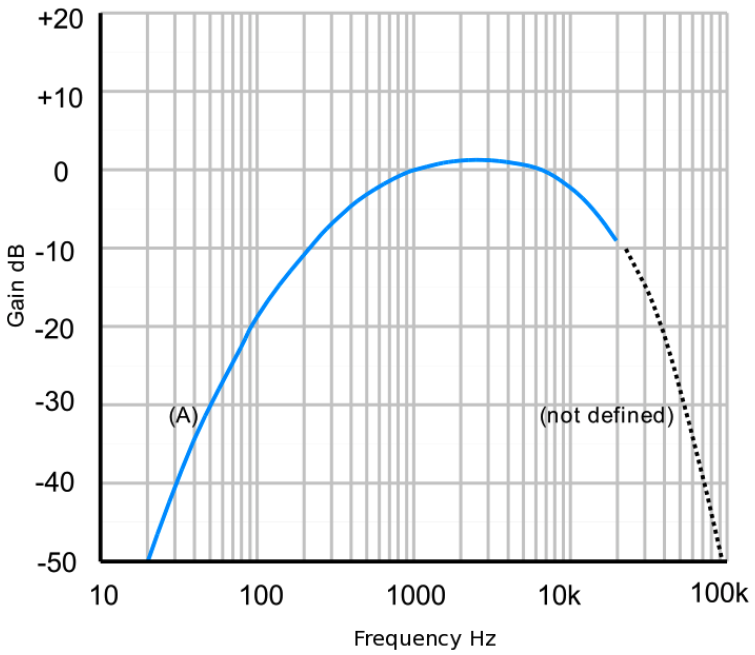


Figure 1. Sound pressure A-weighting scheme according to IEC 61672:2003.

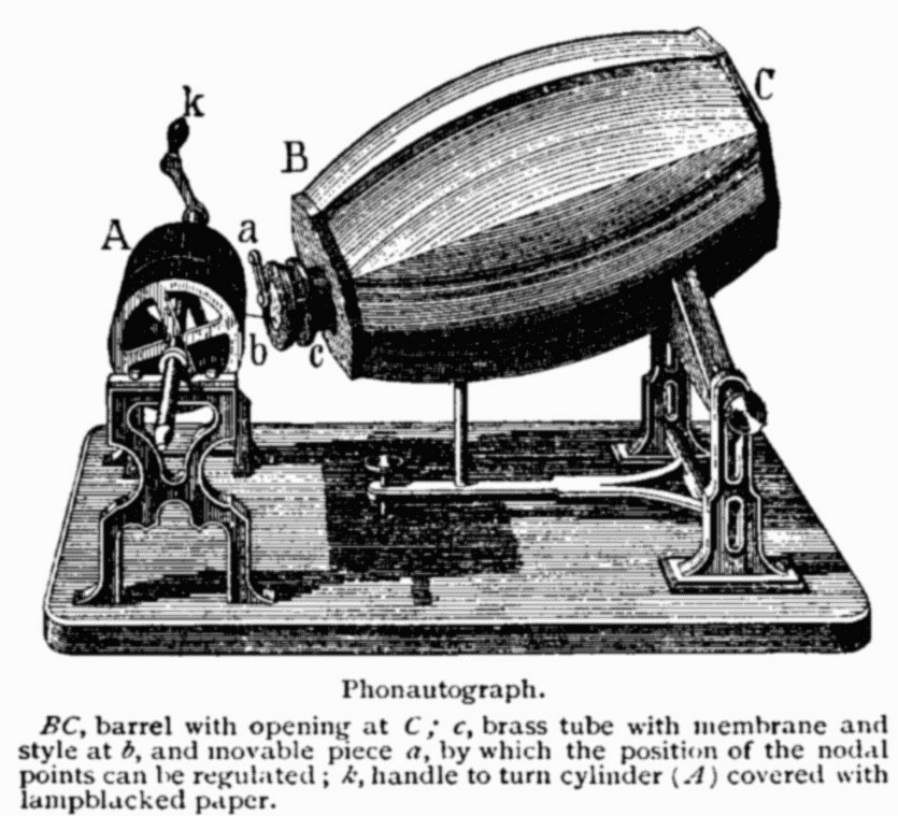
Sound pressure levels weighted by the A scheme are usually labelled as dBA or dB(A). Please note that dB and dBA, like a percent symbol '%', define ratios and not physical units of measurement. A value of 10 dB can refer to completely different sound pressure levels depending on the reference. Also there are no physical units associated with dB.

Observed Properties of Sound

As explained in the previous paragraph, sound is a pressure wave traveling through a medium. In practice, sounds are not exclusively traveling in a homogenous medium from a source to exhaustion. The environment is filled with objects, sometimes sounds are produced in a closed room, and sounds pressure waves may collide with other sounds. The resulting effects of these conditions play a large role when designing multimedia systems. Also, the effects on sound are more significant than on light waves. The three most important ones are echo, reverberation, and interference.

An echo is a reflection of sound, arriving at the listener some time after the original sound. Typical examples are the echo produced by the bottom of a well, by a building, or by the walls of an enclosed room. Sounds is very easily reflected by most materials so echos are always present in every environment. A true echo is a single reflection of the sound source. Mostly, however, many echoes form reverberation. The time delay is the extra distance divided by the speed of sound. When dealing with audible frequencies, the human ear cannot distinguish an echo from the original

sound if the delay is less than 1/10 of a second. Thus, since the velocity of sound is approximately 343 m/s at a normal room temperature of about 20°C, the reflecting object must be more than 16.2 m from the sound source at this temperature for an echo to be heard by a person at the source. For a sound wave to travel that far back and forth it has to have sufficient energy. Normal conversation is usually below this energy threshold.



Phonautograph.

BC, barrel with opening at *C*; *c*, brass tube with membrane and style at *b*, and movable piece *a*, by which the position of the nodal points can be regulated; *k*, handle to turn cylinder (*A*) covered with lampblacked paper.

Figure 2. Phonautograph by Édouard-Léon Scott de Martinville (source: Uncredited 19th century engraving, Wikimedia Commons)

Reverberation is the persistence of sound in a particular space after the original sound wave is exhausted. A reverberation is created when a sound is produced in an enclosed space causing a large number of echoes to build up and then slowly decay as the sound is absorbed by the environment. This is most noticeable to the human ear when the sound source stops but the reflections continue, decreasing in amplitude, until they can no longer be heard. Unless, a room and recording equipment is specially designed to not cause reverberation, reverberation is always present. Even when not audible, multimedia content analysis technique often suffer from not taking into account reverberation. Reverberation is also present during the production of speech in the vocal tract. Reverberation is characterized by the reverberation time. This is the length of this sound decay. It receives special consideration in the architectural design of large chambers, which need to have specific reverberation times to achieve optimum performance for their intended activity.

Interference is the superposition of two or more (sound pressure) waves that results in a new wave pattern. Interference usually refers to the interaction of waves that are correlated or coherent with each other, either because they come from the same source or because they have the same or nearly the same frequency. Interference of sounds causes different effects that are described in wave propagation equations in physics. When designing multimedia systems it is important to know the existence of interference. Also it can be used constructively and destructively. Consider two waves that are in phase, with amplitudes A_1 and A_2. Their troughs and peaks line up and the resultant wave will have amplitude $A = A_1 + A_2$. This is known as constructive interference. If the two waves are 180° out of phase, then one wave's crests will coincide with another wave's troughs and so will tend to cancel out. The resultant amplitude is $A = |A_1 - A_2|$. If $A_1 = A_2$, the resultant amplitude will be zero. This is known as destructive interference. Destructive interference is often use to eliminate unwished sounds, e.g. in noise-canceling earphones. Please note, that even though often mistakenly exchanged, interference is different from masking effects, which are caused by the perceptual properties of the human brain (more about this later).

Recording of Sound

With sound being ubiquitouos and, more importantly, because of mankind’s abilities to use sounds for communication as well as for entertainment purposes, the ability to accurately record it has been a dream for thousands of years. The first device that could record sound mechanically (but could not play it back) was the phonautograph, developed in 1857 by Parisian inventor Édouard-Léon Scott de Martinville. The earliest known recordings of the human voice were phonautograms also made in 1857. These earliest known recordings include a dramatic reading in French of Shakespeare's Othello and music played on a guitar and trumpet. Figure 2 shows a schematic of the device from the inventor’s original records. A barrel with an opening would capture the sound waves and focus them onto a membrane to which a hog's bristle was attached, causing the bristle to move and enabling it to inscribe the sound onto a visual medium. Even though this device was more an early oscillograph than a sound recording device, the concept of the first practical sound recording and reproduction device wasn’t too different. The mechanical phonograph cylinder was invented by Thomas A. Edison in 1877 and patented in 1878. The recordings were initially stored on the outside surface of a strip of tinfoil wrapped around a rotating metal cylinder. This way, play back was possible by using a needle that did not apply as much pressure as in the recording but would convert the mechanical engravings on the cylinder into sound waves that would be mechanically amplified. Figure 3 shows a US postage stamp featuring the device.

Figure 3. Edison's Phonograph on US stamp.

Not surprisingly today's sound recording still obeys the same principles with two main exceptions: First, the sound waves are converted to electrical waves by a microphone and second, most of today's storage media is digital, i.e. sound waves are converted into binary numbers before they are imprinted on the medium. The media themselves, such as CDROM or DAT are a bit more sophisticated than Edison's cylinders. Having said that, we are currently observing the replacement of all of these specialized media with generic media, such as harddisks and flash memory. We therefore decided not to explain the technical details of these, the reader is referred to the bibliography for further information. The next paragraphs, however, will explain the governing principles of modern sound processing.

Microphones

A microphone is an acoustic sensor that converts sound into an electrical signal. The general principle is that sound pressure is inflicted on a membrane which varies it's electrical resistance according to the movement. Most microphones in use today for audio use electromagnetic induction (dynamic microphone) by letting the membrane swing a magnetic field produced by a coil, capacitance change (condenser microphone) by letting the membrane be part of a capacitor which varies capacity with movement, or piezoelectric generation (piezo crystals emit electricity when under pressure). Some modern microphones use light modulation to produce the electric signal by "watching" the mechanical vibration (laser microphones). A single dynamic membrane will not respond linearly to all audio frequencies. Some microphones for this reason utilize multiple membranes for the different parts of the audio spectrum and then combine the resulting signals. The different microphone types have different electrical properties. A complete

microphone also includes a housing and a means of bringing the signal from the element to other equipment (eg. wires or RF capability). These and other characteristics, such as diaphragm size, intended use or orientation of the principal sound input to the principal axis (end- or side-address) of the microphone determine the properties of the recorded sound space. When planning a recording, it is therefore best to survey the current available market and read vendor specifications. The most important characteristics of a microphone is its directionality.

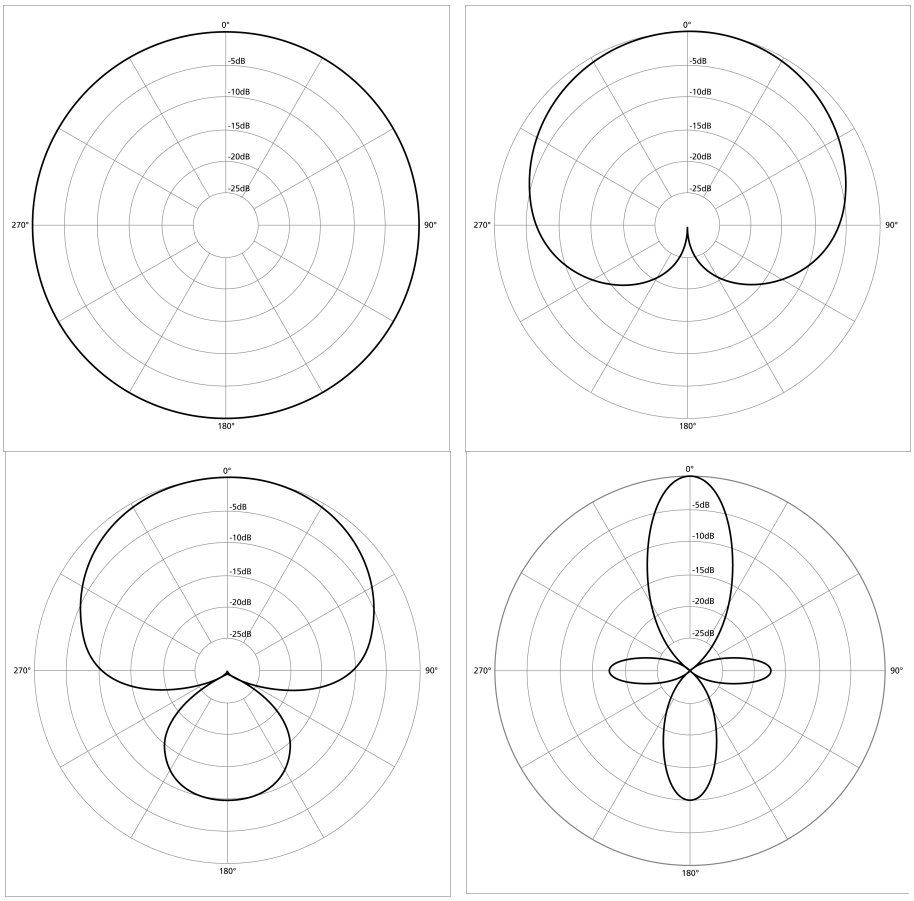


Figure 4: Four common polar patterns of microphones. From left to right: Omnidirectional, cardiod, supercardiod, shotgun (images from Wikimedia Commons).

A microphone's directionality indicates how sensitive it is to sounds arriving at different angles about its central axis. The directionality of a microphone is usually visualized using a polar pattern. Polar patterns represent the location of points that produce the same signal level output in the microphone if a constant sound pressure level is generated from that point. Figure 4 shows some idealized example patterns. The patterns are considered idealized because in the real world, polar patterns are a function of frequency. Manufacturer’s diagrams therefore usually include multiple plots at different frequencies Also, while an omnidirectional microphone's response is generally considered to be a perfect sphere in three dimensions. In the real world, this is not the case. The

body of the microphone is not infinitely small and, as a consequence, it tends to get “in its own way” with respect to sound arriving from the rear, causing a slight flattening of the polar response. This flattening increases as the diameter of the microphone (assuming it's cylindrical) reaches the wavelength of the frequency in question. Therefore, the smallest diameter microphone will give the best omnidirectional characteristics at high frequencies.

Different microphone properties result in different applications: Headset and lavalier microphones are made for hands-free operation. These are small microphones directly worn on the body. Originally, they were held in place with a lanyard worn around the neck, but more often they are fastened to clothing with a clip, pin, tape or magnet. These are directed microphones that allow mobile use for voice recording. These microphones are in everyday use for video conferencing, personal recording, and dictation applications.

A parabolic microphone uses a parabolic reflector to collect and focus sound waves onto a microphone receiver, very similar to a satellite dish. Typical uses of this microphone, which has unusually focused front sensitivity and can pick up sounds from many meters away, include nature recording, outdoor sporting events, and eavesdropping. These microphones are not typically used for standard recording applications, because they tend to have poor low-frequency response as a side effect of their design. However, machine intelligence may be able to infer information from them (e.g. in connection with a surveillance camera).

Noise-canceling microphones have a highly directional design intended for noisy environments when direct attachment to the body is not desirable. One use is on loud concert stages for vocalists. Often, noise-canceling microphones combine signals received from two membranes that are in opposite electrical polarity or are processed electronically later. The main membrane is mounted closest to the intended source and the second is positioned farther away from the source so that it can pick up environmental sounds to be subtracted from the main signal by destructive interference.

Arrays of omnidirectional microphones are best to pick up as much sound from the environment as possible. These are usually used for auditory scene analysis where objects can be located by analyzing the time delay of arrival between different microphones (due to speed of sound). Also the signal quality can be enhanced by combining the signals from a larger set of microphones. This technique is often used in speech recognition, when head or body-mounted microphones are not desirable.

The output of a microphone is usually amplified using an analog amplifier before being digitized. Some microphones already output a digital signal directly as standardized by the AES 42 standard. The next paragraph will explain digitization.

Digitization of Sound

The electric current output by a microphone, and maybe amplified and/or mixed by further equipment is a continuous electrical signal, with the voltage directly proportional to the sound pressure. In practice, sound recording has a linear are for certain sound pressure levels and frequency ranges and non-linear areas if the sound pressure and/or the captured frequency is too low or too high. If the sound pressure level is too low, the signal will mostly just be zero if it is too high, it will reach an internal clipping point (which in the worst case is a short-circuit) and will be severely distorted. Even if it does not reach the clipping point, the non-linear behavior of sound processing devices will lead to distortion when the captured signal is outside the non-linear scope. This is referred to as the signal being overdriven. When the signal is outside the linear frequency range of the recording device, harmonic distortion will be introduced. For example, a nicely shaped sinus curve might be converted into anything not nicely shaped anymore. Figure 5 shows an example.

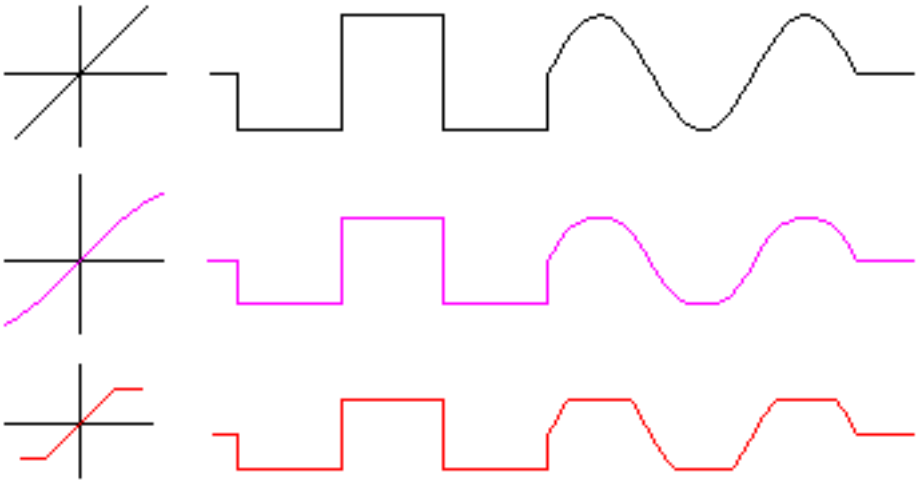


Figure 5. The diagrams on the left show the amplification behavior, the curves on the right show the results for different input signal shapes. Top: Linear behavior, original signal, second row: typical analog behavior and distorted signal due to overdrive, bottom: typical digitization behavior and distorted signal from clipping.

The main problem is that, even if standardized, every sound processing device has slightly different linear ranges. Also, sound cables, especially when very long, might inhibit certain frequencies and since they often work as “involuntary antennas” might introduce electric distortion from the outside, the most current one being a “buzz” from the 50Hz/60Hz electrical system. Then, recording media, such as the old venyl records or audio cassettes introduce their own non-linearities and the effects stack with every copy made. Therefore, in the last two decades, the sound processing has shifted from analog to digital. At the time of writing this book, many microphones,

mixers, and pre-amplifiers are still analog but the storage and processing is digital. With standards, like the aforementioned AES 42 getting more and more popular, digitization will become a much earlier part of the processing chain soon.

Digitizing is the representation of a signal by a discrete set of its samples. Instead of representing the sound signal by an electrical current proportional to its sound pressure, the signal is represented by on-off patterns that represent sample values of the analog signal at certain fixed points. The on-off patterns are much less susceptible to the distortions outlined above, especially copying is usually lossless. Conceptually, digitization works in two parts, illustrated in Figure 6.
- Discretization: The analog signal is read at regular time intervals (sampling rate), sampling the value of the signal at that point in time. One such reading is called a sample.
- Quantization: Samples are rounded to a fixed set of numbers (such as integers), a process known as quantization.

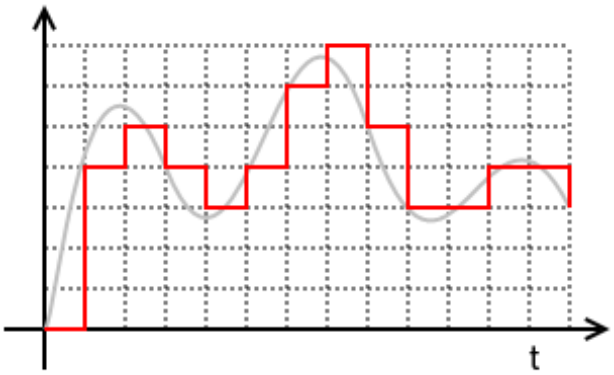


Figure 6. Digital representation of an analog signal. Both amplitude and time axis are discretized.

A series of quantized samples can be transformed back into an analog output that approximates the original analog representation by generating the signal represented by each sample. The sampling rate and the number of bits used to represent the sample values determine how close such an approximation to the analog signal a digitization will be.

The error introduced by the quantization is called quantization noise and affects how accurately the amplitude can be represented. Very few bits for the samples will result in the signal only being represented coarsely and will affect the perceived dynamic of the sound as well as introduce high-frequency artifacts. Typical bit representations for audio are 8, 16, and 24 bits.

The error introduced by the sampling rate is called discretization error and determines the maximum frequency that can be represented in the signal. This upper frequency limit is determined by the so-called Nyquist frequency. The Nyquist frequency, named after the Swedish-American engineer Harry Nyquist or the Nyquist–Shannon sampling theorem, is half the sampling

frequency of a discrete signal processing system. In other words, if a function *x(t)* contains no frequencies higher than *B* hertz, it is completely determined by giving its ordinates at a series of points spaced *1/(2B)* seconds apart.

The proof of this fundamental theorem can be found in the research papers at the end of this chapter. In this overview chapter we will stick with an illustrating example.

To illustrate the necessity of $fs > 2B$, consider the sinusoid:

$$x(t) = \cos(2\pi Bt + \theta) \equiv \cos(2\pi Bt)\cos(\theta) - \sin(2\pi Bt)\sin(\theta).$$

With $fs = 2B$ or equivalently $T = 1/(2B)$, the samples are given by:

$$x(nT) = \cos(\pi n)\cos(\theta) - \underbrace{\sin(\pi n)}_{0}\sin(\theta) = \cos(\pi n)\cos(\theta).$$

Those samples cannot be distinguished from the samples of:

$$y(t) = \cos(2\pi Bt)\cos(\theta).$$

But for any θ such that $sin(\theta) \neq 0$, *x(t)* and *y(t)* have different amplitudes and different phase. Figure 7 illustrates this further.

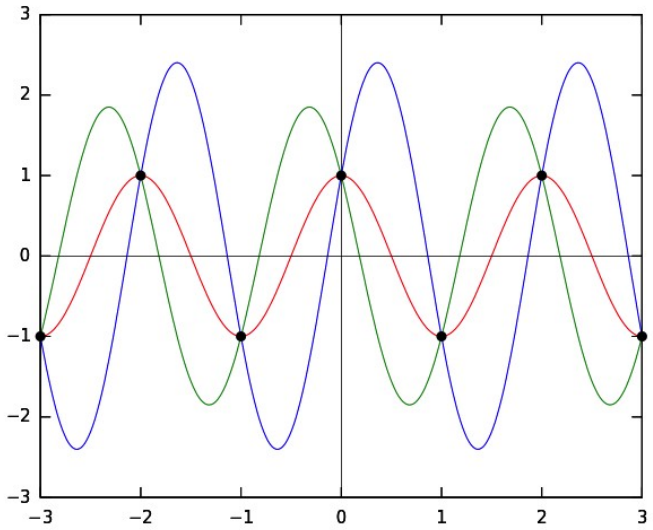


Figure 7. Three possible analog signals for the same sampling points.

The Nyquist theorem is not at all limited to sound signals, it is true for the digitization of any signal. However, we discuss it here, since, from all multimedia formats sound formats are most influenced by this limit. Since the maximum frequency perceptible by the human auditory system is about 22 kHz, compact discs sample at 44Khz. Human speech which usually peaks at about 6-8kHz is considered completely represented by a 16kHz sampling frequency. Sampling frequencies

about 44kHz, such as 48kHz and 96kHz are frequently used by professional audio recording equipment. If the analog equipment supports it, these devices are able to capture frequencies that are not able to be perceived by the human ear. However, it allows for better reproduction of overtones and further processing, such as digital filters and machine learning, might use the higher frequencies too.

Reproduction of Sound

So far we assumed the existence of a sound signal. On any place on earth this assumption is a safe one, since sound pressure levels can be virtually measured everywhere. However, if sound is to be reproduced from a storage, one has to invent special devices to do so. The most often used device for sound reproduction is the loudspeaker.

The loudspeaker is the exact reverse of a microphone. A typical experiment for electronic experimental kits sold to teenagers is, to reverse them because simple microphones can serve as loudspeakers and loudspeakers can also act as microphones. Again, a loudspeaker (or speaker) is an electroacoustic transducer that converts an electrical signal into sound. The speaker pulses in correspondence with the variations of an electrical signal and causes sound waves to propagate through a medium such as air or water. Loudspeakers usually consists of a membrane that is driven back and forth and made to oscillate using an electricity to mechanical force converter, such as an electro magnet or a piezo crystal. This core part is usually called driver. The term loudspeaker therefore can refer to individual drivers or to an integrated system of drivers in an enclosure. The role of the enclosure, apart from providing a place to mount the drivers, is to prevent sound waves emanating from the back of a driver from interfering destructively, i.e. by causing cancellation, with those from the front. To adequately reproduce a wide range of frequencies, most loudspeakers require a combination of drivers with different properties. Each individual drivers is then used to reproduce a different frequency range. Common driver types include subwoofers (very low frequencies, typically below 120 Hz); woofers (low frequencies); mid-range speakers (middle frequencies); tweeters (high frequencies); and sometimes supertweeters, which are optimized for the highest audible frequencies. When multiple drivers are used in a system, a network of electrical filters, called a crossover, is used to separate the incoming signal into different frequency ranges and to route them to the appropriate driver. A loudspeaker system with *n* separate frequency bands is called *n*-way speakers. Typical home audio devices have a 3-way speaker system, consisting of a woofer, a mid-range speaker, and a tweeter. Like microphones, loudspeakers have a directionality, i.e. their frequency (re-)production properties varies in space. Figure 8 shows the directionality of a typical column-shaper home audio system speaker.

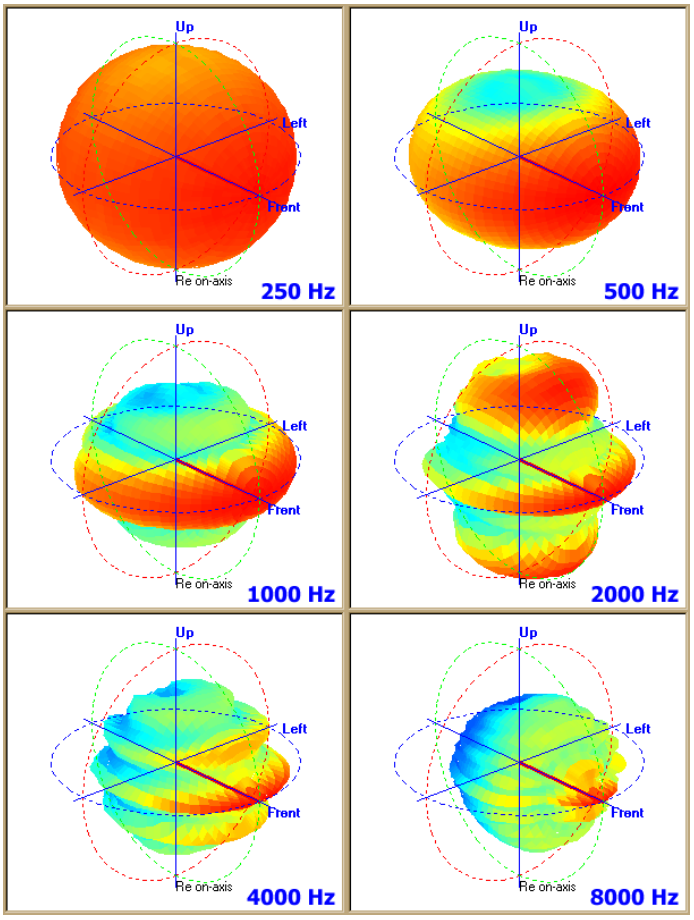


Figure 8. Polar patterns of a typical home speaker system that consists of four drivers at different frequencies (source: Wikimedia Commons).

Needless to say speakers are designed with different directionality for different applications, e.g. car speakers have a different polar pattern than the speakers used to making announcements in a supermarket. Other factors that determine the properties of a loudspeaker are the rated power, which determines the maximum input a speaker can take before being destroyed, the maximum sound pressure level (SPL), which defines how much sound pressure can be emitted by the speaker, the impedance, which determines the electrical compatibility with different amplifiers, the crossover frequencies, which define the nominal frequency boundaries of the signal devision by the drivers, and -- last but not least, the frequency range which determines the linear frequency response range of the speaker system. The enclosure type, eg. sealed or bass reflex, determines some of the perceptual properties of the loudspeakers. Another important factor for the quality of the sound reproduction is the relationship between the number of channels used (e.g., two four, or six), the way they have been encoded (eg. stereo or surround), and the way the speakers are placed in the room when reproducing sound.

Discussing loudspeakers in detail would require another whole series of books. Their design is a profession and research is still being performed in this area -- although, of course, not as part of computer science. The reader is therefore referred to the bibliography. It is very important to remember though that loudspeakers and microphones are the most variable elements in terms of perceived quality. Thus, apart from lossy compression, they are usually responsible for most distortion and audible differences when comparing sound systems. A practical advice to the reader is: Whenever experimenting with sound algorithms use high-quality headphones.

Production of Speech

As already discussed in the previous paragraph, except for a vacuum, sound is everywhere, even when not audible. The production of random noise is therefore relatively easy. However, modulating sound in a way suitable for communication requires a sophisticated apparatus. Even though humans are not the only species that can produce sophisticated sounds, compare e.g. parrots which have an even greater ability to reproduce sounds from their environment, humans seem to have developed the most sophisticated expressiveness. Speech is therefore the most important communication means for human beings. The following section will introduce the basic anatomy and properties of the human speech generation apparatus as well as a few external factors that influence speech production.

Let’s start with some facts. The frequency range of speech is between 80Hz and about 5kHz. The pitch of the human voice is between 120 and 160Hz for adult males and between 220 and 330 Hz for women and children. Vowels can reach frequencies up to about 5kHz. The highest, frequencies are emitted by sibilants. Their frequencies can easily reach into the non-audible spectrum (above 20kHz). The dynamics of speech is relatively high compared to many other sounds sources, such as some musical instruments (see below). In general the volume of human voice is limited to the sound energy the human body can produce. In a 60cm distance from the mouth, the typical sound volume of the human voice is about 60dBA. A stronger voice can raise the volume by about 6dB. Yelling measures about 76dBA (males) and 68dBA (females).

The research field that investigates how sounds are produced is part of linguistics and is called articulatory phonetics. The field researches how the tongue, lips, jaw, and other speech organs are involved in making a speech sound make contact. Almost all organs have additional functions and are not exclusively used for speech production. Also, of course, different organs have more than one function in speech production, and, to make things even more complex, the same sounds can be produced by different combinations of organs. Figure 9 shows a schematic drawing of the human speech production apparatus.

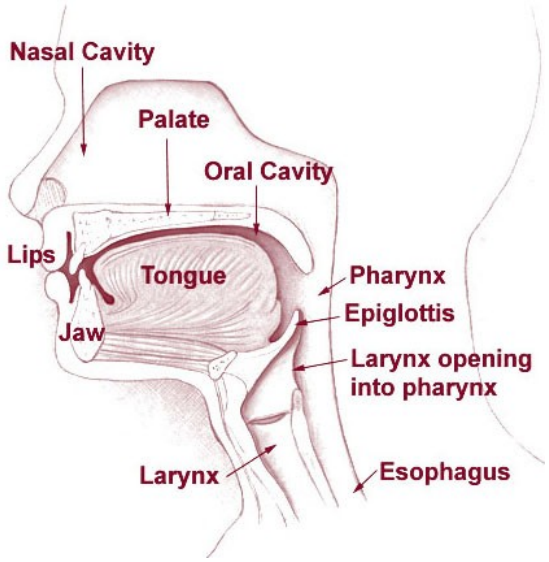


Figure 9. A schematic of the human speech production apparatus. (Source: National Cancer Institute)

Sound is pressure waves traveling through air, Therefore, human speech is directly connected to the body's respiratory system. The mechanism for generating the speech can be mainly subdivided into three parts: the lungs, the vocal folds within the larynx, and the articulators. The lung must produce adequate airflow and air pressure to vibrate vocal folds. The vocal folds (inside the larynx) are a vibrating valve that chops up the airflow from the lungs into audible pulses that form the so-called laryngeal sound source. The muscles of the larynx adjust the length and tension of the vocal folds to adjust pitch and tone. The articulators (the parts of the vocal tract above the larynx consisting of tongue, palate, jaw, lips, etc.) articulate and filter the sound emanating from the larynx and to some degree can interact with the laryngeal airflow to strengthen it or weaken it as a sound source. The mechanism described until now allows to generate vowels. In order to generate consonants two opposing organs in the speech production system have to make contact. The contact point is called place of articulation of a certain consonant. Linguists say, an obstruction occurs in the vocal tract between a moving articulator (typically some part of the tongue) and a stationary articulator (typically some part of the roof of the mouth). Speech sounds are usually classified as stop consonants (with blocked airflow, eg. English 'p', 't', or 'k'), fricative consonants (with partially blocked and therefore strongly turbulent airflow, eg. English `f` or `v`), approximants (with only slight turbulence, eg. English `w` and `r`), and vowels (with full unimpeded airflow, e.g. English 'a', 'e','i', and 'o'). A hybrid class is the so-called Affricates; they are sequences of stop plus fricative (e.g. English `ch` or `j`). Fricatives are usually further subdivided into sibilants and lateral fricatives. The former are a type of fricative where the airflow is guided by a groove in the tongue toward the teeth, creating a high-pitched and very distinctive sound (e.g. English `s` or `z'). The latter are a rare type of fricative (non-existent in English),

where the frication occurs on one or both sides of the edge of the tongue. Other classes exist and there is a variation in different languages. Vowels are usually classified into monophtongs, having a single vowel quality, and diphtongs, vowels which manifest a clear change in quality from start to end as in the words *bite, bate, or boat.*

Consonants and vowels are similar to something like the building blocks of speech. Linguists refer to these building blocks as phonemes. American English has 41 phoneme, although the number varies according to the dialect of the speaker and the system of the linguist doing the classification. The concrete pronunciation of a phoneme is dependent on the previously and the following uttered speech sounds. It also depends on the type of speech (e.g. whispering vs screaming), emotional state of the speaker, as well as anatomy of the throat, age, native language and dialect, and social and environmental surrounding. Diseases of the lungs or the vocal cords, including paralysis, respiratory infections, vocal fold nodules and cancers of the lungs and throat affect the sound and clearness of speech, diseases and disorders of the brain, including speech processing disorders, where impaired motor planning, nerve transmission, phonological processing or perception of the message (as opposed to the actual sound) leads to poor speech production and usually affects the speed measured in syllables per minute. Also, hearing problems can lead to phonological problems. Those who are hard of hearing or deaf may be considered to fall into this category. Articulatory problems, such as stuttering, lisping, cleft palate, or nerve damage leading to problems in articulation. In other words, the actual frequency pattern of a specific uttered consonant or vowel is underlies large variance.

Environmental effects and the brain processing input from other modalities, such as sight or touch, can affect speech greatly. This so-called Lombard effect describes an involuntary tendency to increase volume, change pitch, or adjust duration and sound of syllables as a response to external noise. This compensation effect results in an increase in the auditory signal-to-noise ratio of the speaker's spoken words. This is a general phenomenon observed in many animals, including birds and whales (see research papers). This is also one reason why automatic speech recognition algorithms trained in a quiet environment are very difficult to transform to noisy environments. For example, an algorithm trained in the cubicle of the developer will hardly work in a car.

A standard way to visualize and further analyze sound patterns is the spectrogram. A spectrogram is an image that shows how the spectral density of a signal varies with time, i.e. it shows the distribution of the energies in different frequency bands in time. Figure 10 shows an example. Reading spectrograms is something computer linguists, for example, learn very early. It is a handy skill and helps getting a feeling for the properties of a speech signal. Today's machine learning algorithms have made this skill not a main priority anymore, however, the vocabulary associated with it is still used for describing the properties of a speech signal.

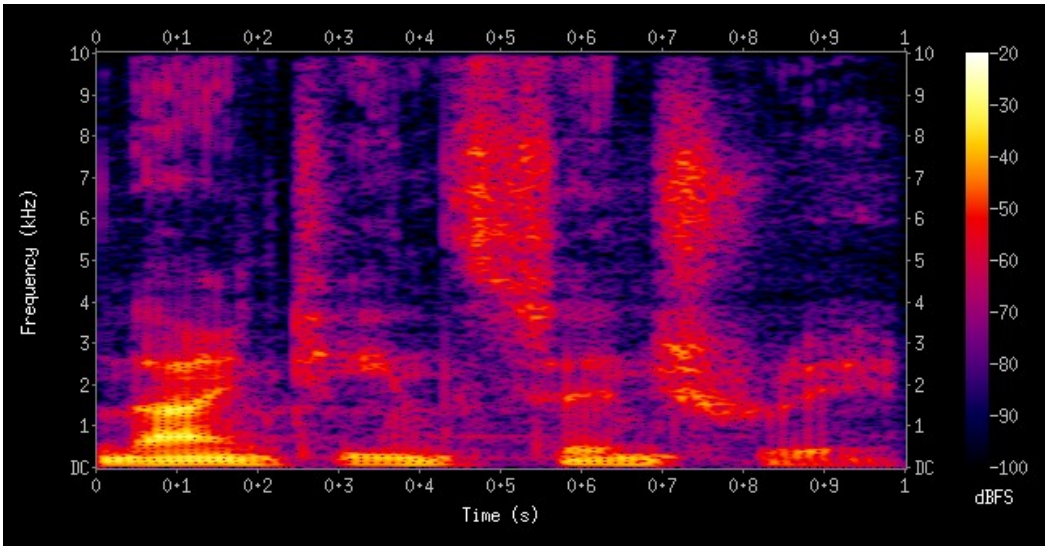


Figure 10. A spectrogram showing a male voice saying: “nineteenth century”. The yellow bands of energy are the formants (source: Wikimedia commons).

Each phoneme is distinguished by its own unique pattern in the spectrogram. For voiced phonemes, the signature involves large concentrations of energy, so-called formants. As outlined above, formant values vary widely from person to person. So spectrogram reading mainly means to learn to recognize patterns which are independent of particular frequencies and which identify the various phonemes with a high degree of reliability. Relatively static formants are found in the monophthong vowels and the nasals; formants which are more variable over time are found in the diphthong vowels and the approximants. The monophthong vowels have the strongest and most stable formants. So these vowels can usually be easily distinguished by the frequency values of the first two or three formants, which are called F1, F2, and F3. Depending on the phoneme, F1 varies from about 300Hz to 1000Hz, F2 from 850Hz to 2500Hz, and F3 from 2300Hz to 3000Hz. Higher formants such as F4 and F5 are not used for communication anymore but are indicative of the speaker’s voice. As a result of the low bandwidth and the Nyquist theorem F4 and F5 are usually lost in telephone speech -- as are many of the speakers’ individual voice characteristics. Unvoiced speech sounds are not usually said to have formants. Still, plosives are usually recognized as a great burst of energy across all frequencies occurring after a short relative silence. Aspirates and fricatives are recognized as “large clouds” of smooth energy along both the time and frequency axes.

As the reader may have noticed from this section, the scientific analysis and description of speech production has keept several research fields busy. Unfortunately, we can only provide a glimpse at this field. For further information, please refer to the bibliography. The next section describes music production, which is even more complex.

Production of Music

Apart from human's natural ability to produce systematic sounds for communication, humans have also developed artificial means of producing sounds from scratch (i.e., without reproducing). Researchers have discovered various archaeological evidence of musical instruments in many parts of the world. Some are as much as 37,000 years old. Of course, only artifacts made from durable materials survive to be found. So the oldest instrument might be even older. What this tell us though is that music plays a very important role in human kind and seems to be connected to their nature. The building and use of musical instruments vary with history and culture as do the sounds that these instruments produce. The concrete sound emitted, of course, also with the musician playing it. In other words, music is a matter of art. However, for multimedia computing we are interested in the describing the general properties of instruments so that we can leverage these for audio compression, detecting instruments, manipulating recordings, or artificially synthesizing them. Let's start with some basic vocabulary.

The fundamental frequency, abbreviated as f_0 or F_0 (speak: f-zero), is the inverse of a period length of a periodic signal. Pitch represents the perceived fundamental frequency of a sound. While the fundamental frequency can be precisely determined through physical measurement, it may differ from the perceived pitch because of overtones. An overtone is either a harmonic or partial (non-harmonic) resonance. In most musical instruments, the frequencies of these tones are close to the harmonics. The harmonic of a wave is a component frequency of the signal that is an integer multiple of the fundamental frequency, eg. when the fundamental frequency is f, the harmonics have frequencies f, 2f, 3f, 4f, etc. The most important property of the harmonics is that they are all periodic at the fundamental frequency. In other words, the sum of the harmonics is also periodic at that frequency. A term very particular to music is timbre. The word timbre is used to describe the quality of sound that distinguishes different types of sound production, such as different musical instruments. The physical characteristics of sound that mediate the perception of timbre include spectrum and time envelope. Spectrum, which means frequency spectrum, would be described by a musician as the sum of distinct frequencies emitted by an instrument playing a particular note with the strongest frequency being the fundamental frequency. In western music, instruments are normally tuned to the orchestral tuning note A = 440 Hz. So when this tuning note is played by an instrument, the emerging sound is actually a combination of frequency components, including 440 Hz, 880 Hz, 1320 Hz, 1760 Hz, etc. (harmonics) and some partials. The balance of the amplitudes of the different spectral components is responsible for the characteristic sound of each instrument. The timbre of a sound is also greatly affected by its time envelope. The model typically used to describe it divides the sound development into four stages: Attack, decay, sustain, and release (or ADSR envelope). Attack defines the time from when the sound is activated to its reaching the full amplitude. Decay defines the time the sound needs for dropping from maximum amplitude to

sustain level. Sustain defines the volume level the sound is at until the note is released. This parameter defines a volume level not a time, as the time to sustain the sound is defined by the musician. Release finally describes the time needed for the sound to fade when the note ends. Psychoacoustics uses the word tone quality and tone color as synonyms for timbre. Pitch, loudness, and timbre are the three major auditory attributes that describe a sound.

The most commonly used category system for musical instruments in use in the western world today divides instruments into string instruments, wind instruments and percussion instruments. A string instrument produces sound by means of vibrating strings. The most common string instruments include banjo, cello, double bass, guitar, harp, mandolin, ukulele, violin, viola, and the piano. The vibration of the strings have the form of standing waves which produce a single fundamental frequency (pitch) and all harmonics of that fundamental frequency simultaneously. These frequencies depend upon the tension, mass and length of the string. The harmonics make the sound timbre fuller and richer than the fundamental alone. The particular mix of harmonics present depends upon the method of excitation of the string. The timbre of the sound produced by the string varies significantly depending on the method of excitation of the string. In a violin the string is bowed and sometimes plucked while in a piano the string is struck by a falling hammer. In addition to the string properties and the method of excitation, the sound timbre is significantly affected by resonances in the body of the instrument itself. A wind instrument contains some type of resonator, usually a tube, in which a column of air is set into vibration by the musician blowing into the end of the resonator. The pitch of the vibration is determined by the length of the tube. The length is usually varied artificially by manual modifications of the effective length of the vibrating column of air, eg. by holes in the tube that are cover or uncovered to create different pitches. The sound wave travels down the tube, reflects at one end and comes back. It then reflects at the other end and starts over again. For a note in the lowest register of the flute, for example, the round trip constitutes one cycle of the vibration. The longer the tube, the longer the time taken for the round trip, and so the lower the frequency. The most common wind instruments include flutes, clarinets, oboes, bassons, saxophones, trumpets. A percussion instrument is produces sound by being hit, shaken, rubbed, scraped, or by any other action which sets it into vibration directly. Instruments in this category include drums, bells, or xylophones. The acoustics of percussion instruments is the most complex as most percussion instruments vibrate in a rather complex ways. In general, at low to medium amplitudes, their vibrations can be conveniently described by the terms introduced in this chapter. At large amplitude, however, they may show distinctly nonlinear or chaotic behavior. Percussion instruments have the highest variance in frequency and amplitude range and are therefore the most difficult to process.

In the end, a musical piece contains a mixture of a variety of instruments, including human voices. Once mixed, separation of each individual instrument is very difficult since it would require an adequate model of the instrument’s behavior in its environment and with the used recording

equipment. For this reason, music is not only recorded and digitized but also saved in a note-like format, called MIDI, that defines a protocol to control electronic instruments. Electronic instruments have long tried to achieve mimic traditional ones through a process called music synthesis, which is briefly described in the next section.

Synthesis of Sound

The last section of this chapter glimpses over a topic that, even though much more recent than music and speech production, could easily fill many book: The artificial generation of speech and music, called synthesis. The first music synthesizers go back to as early as 1876. Back then, as is today, the main goal was not necessarily to imitate a physical music instrument correctly. Often the goal was to create new sounds of artistic value. When it comes to the exact simulation of a real instrument, the difficulty and complexity of the task depends of course on the properties of the physical instrument. Simulating a simple flute is much easier than a piano or a particular unique organ. It's not unusual that particular algorithms are invented for a particular subtype of instrument. In general though, modern music synthesis is performed by physical modeling of the instrument as well as incorporating original samples of the instrument, so-called wavetables. After a long period of time and a variety of proposed techniques, research ihas currently converged to doing the same for speech synthesis: Synthesized speech is often created by concatenating pieces of recorded speech from a database. Systems currently differ in the size of the stored speech units. a system that stores phones or tuples of phones (so-called diphones) provides the largest output range but may lack clarity and naturality of the voice output. Trading of output range for usage in a specific domain the storage of entire words or even sentences allows for higher quality output. The database is usually combined with a model of the vocal tract (such as LPC, see chapter XXX) and other human voice characteristics to create a completely synthetic voice output. This concept of adaptive concatenative sound synthesis is the same as for both speech and music synthesis. The many products and research projects apply this concept in many different ways, which are impossible to describe in brevity here. The reader might refer to the bibliography in the field (and at the end of this chapter).

Exercises

1. How many dB are 50%, 1%, 0.01%, and 200%? How many dB can be stored in 16bits, 24bits, and 32bits?
2. List the factors that would influence echo and reverberation in a lecture hall.
3. You are a researcher working on a project that has to do sound recordings frequently. Unfortunately, you are forced to share your office with a room mate who needs a very noisy server

farm right standing beside him. Given no social rules or limitations: What would be the best thing to do to isolate the noise?

4. Discuss what would be the best directionality for a microphone that is used for field studies where you interview people in noisy environments.
5. Discuss and experiment with ideas to reconstruct frequencies beyond the Nyquist limit. What are the trade-offs?
6. Explain how the Nyquist limit sometime becomes visible in the image and video domain. What are the typical artifacts?
7. Explain the artifacts you would expect from a microphone/loudspeaker that is forced to record/play sound a) outside its frequency range b) outside its amplitude range.
8. Assume you would like to record a seminar with many participants in a classroom. What environmental noise would you expect?
9. When a signal received by a microphone is amplified and passed out of a loudspeaker. The sound from the loudspeaker might be received by the microphone again, amplified further, and then passed out through the loudspeaker again. The effect is known as Larsen effect or more colloquial as feedback loop. Describe what happens and how the signal looks like.
10. Which differences would you expect to see in a spectrogram between male and female speakers? What about the difference between younger and older speakers?
11. What would be the typical spectrogram of a flute, a violin, or a drum?
12. Implement an ADSR envelop filter and play around with it. Apply it to different sounds and waveforms, including noise.

Literature

- Data Interchange on Read-only 120 mm Optical Data Disks (CD-ROM). ECMA-130. June 1996.
- F. Alton Everst, Ken Pohlmann: Master Handbook of Acoustics, McGraw-Hill, 5th edition, June 2009.
- D. T. Blackstock: Fundamentals of Physical Acoustics, Wiley-Interscience, 1st edition, February 2000.
- B. Gold, N. Morgan: Speech And Audio Signal Processing: Processing And Perception Of Speech And Music, Wiley, 1st edition, August 2006.
- A. C. Bickford, R. Floyd: Articulatory Phonetics: Tools for Analyzing the World's Languages, 4th edition, SIL International, July 2006.
- T. Kientzle: A Programmer's Guide to Sound, Addison-Wesley, October 1997.

Web Links

- MIDI specifications: http://www.midi.org/
- Neck anatomy: http://training.seer.cancer.gov/head-neck/anatomy/overview.html

Research Papers

- H. Nyquist. "Certain topics in telegraph transmission theory", Trans. AIEE, vol. 47, pp. 617-644, Apr. 1928. Reprint available as classic paper in: Proceedings of the IEEE, Vol. 90, No. 2, Feb 2002.
- É. Lombard. "Le signe de l'élévation de la voix", Annales des Maladies de L'Oreille et du Larynx, Vol. XXXVII, No. 2, pp. 101–119, 1911.
- Slabbekoorn H, Peet M. "Birds sing at a higher pitch in urban noise". Nature. 424(6946):267, 2003.
- Junqua JC. "The Lombard reflex and its role on human listeners and automatic speech recognizers", Journal of the Acoustic Society of America, Jan;93(1):510-24, 1993.
- Scheifele PM, Andrew S, Cooper RA, Darre M, Musiek FE, Max L. "St. Lawrence River beluga Indication of a Lombard vocal response in the St. Lawrence River Beluga". Journal of the Acoustic Society of America. 117(3 Pt 1):1486-92, 2005.